Journal of Computer Science and Cybernetics, V.41, N.1 (2025), 79-94
DOI no 10.15625/1813-9663/20607

A TWO-STAGE FRAMEWORK FOR VIETNAMESE COMPARATIVE OPINION QUINTUPLE EXTRACTION

DANG VAN THIN*, NGUYEN THI THUY, DUONG NGOC HAO, NGAN LUU-THUY NGUYEN

VNU-HCM University of Information Technology, Quarter 6, Linh Trung Ward, Thu Duc City, Ho Chi Minh City, Viet Nam

Abstract. Comparative opinion mining is an important subtask of opinion mining. It aims to identify comparative reviews and extract the comparative elements in quintuples. This task, known as Comparative Opinion Quintuple Extraction (COQE), has two main sub-tasks: Comparative Sentence Identification (CSI) and Comparative Element Extraction (CEE). In this paper, we introduce an effective two-stage framework for the COQE task specifically designed for the Vietnamese language. The first stage leverages the power of fine-tuning different BERT-based language models to identify the comparative sentences. We then formulate the comparison extraction task as a conditional text generation problem and apply a multi-task instruction prompting architecture based on generative language models. Furthermore, we also employ a data augmentation technique to increase the training data samples. Our experimental results on the VCOM dataset [1] show that our framework outperforms existing methods and achieves state-of-the-art performance on the test set. We also conduct a detailed analysis to provide insights for future research on this topic.

Keywords. Comparative mining, Vietnamese language, two stage frameworks, multi-task prompting tuning.

1. INTRODUCTION

Opinion Mining or Sentiment Analysis is one of the active topics of Natural Language Processing (NLP) that analyzes the attitudes or emotions of people towards specific entities [2]. The opinion is divided into two different types: regular opinion and comparative opinion. A regular opinion can be direct or indirect, expressing a judgment or sentiment on an entity or aspect of the entity. Meanwhile a comparative opinion expresses a relation of similarities or differences between entities in comparison form (e.g. "A is better than B"). For business organizations, the comparative opinions of users play a very important role in determining competitor information compared to their own products and services.

To the best of our knowledge, the work of Jindal and Bing [3] is considered a pioneering study in the analysis of comparative opinions proposing two sub-tasks, called Comparative Sentence Identification and Comparative Element Extraction. Recently, Liu et al. [4] introduced a new task called Comparative Opinion Quintuple Extraction (COQE). This task aims

*Corresponding author.
E-mail addresses: thindv@gm.uit.edu.vn (D.V. Thin); 21521514@gm.uit.edu.vn (N.T. Thuy); haodn@uit.edu.vn (D.N. Hao); ngannlt@uit.edu.vn (N.L.T. Ngan).

to identify comparative sentences and extract the elements within the comparative quintuples. This work has motivated recent research on the COQE problem where many different methods [5, 6, 7] have been proposed. For the Vietnamese language, Le et al. [1] released a benchmark dataset at the VLSP 2023 ComOM shared task. Unlike other published datasets [4] in English, this dataset requires approaches to extract the corresponding index positions of elements within a sentence.

In this paper, we introduce a two-stage framework, which achieved the best performance on the ComOM shared task [1]. Our framework addresses the COQE task in two stages: first, identifying comparative reviews and then extracting the elements within the quintuples from those reviews. The main contributions of the paper are as follows: (1) we present a two-stage framework for the Vietnamese COQE task that utilizes the power of BERTology for identify the comparative sentence and generative language models to extract the comparative elements in the quintuplets; (2) the experimental results show that our framework outperforms previous approaches, establishing a new state-of-the-art on the VCOM dataset [1]; (3) we conduct a detailed error analysis of our method providing insights for future research.

The structure of this paper is as follows: Section 2 reviews the related work on comparative opinion quintuple extraction. Section 3 details our proposed methodology, followed by the experimental settings in Section 4. The experimental results and discussion are provided in Section 5. Finally, Section 6 concludes the paper with a summary of our findings.

2. RELATED WORK

Comparative Opinion Quintuple Extraction was first introduced by Liu et al. [4], and aims to identify comparative sentences from product reviews and extract all comparative opinion quintuples: including Subject, Object, Comparative Aspect, Comparative Opinion, and Comparative Preference. The concept of comparative opinion mining was first proposed by Jinda and Liu [8] with two tasks. For the end-to-end comparative opinion task, Liu et al. [4] proposed a three-stage neural network approach to identify the comparative reviews and extract the comparison quintuplets. In the first stage, the authors introduced a multi-task learning framework based on BERT-based models combined with the CRF model to classify the comparative sentences and four first elements simultaneously. Xu et al. [9] presented a BERT-based model combined with a Graph Convolutional Network to improve the performance of the end-to-end model. Another work of [5] proposed an end-to-end approach called UniCOQE by transforming this task into the text generation problem. The UniCOQE architecture reduced the order bias of generative models during training by combining a generative paradigm with a set-matching strategy based on the Hungarian algorithm. Due to the limitation of the training dataset, Xu et al. [6] proposed a data augmentation method using ChatGPT.

In Vietnamese, Bach et al. [10] conducted one of the first studies addressing the tasks of identifying comparative sentences and recognizing relationships in review data. More recently, the ComOM shared task [1] was introduced to encourage researchers to develop and evaluate methods for comparative mining. Several approaches have been proposed, including BERT-LSTM-CRF, BERT2BERT [1], the Multi-Stage Framework [4], and Three-Stage BERTs [11]. These methods often break the task into smaller sub-tasks, solving

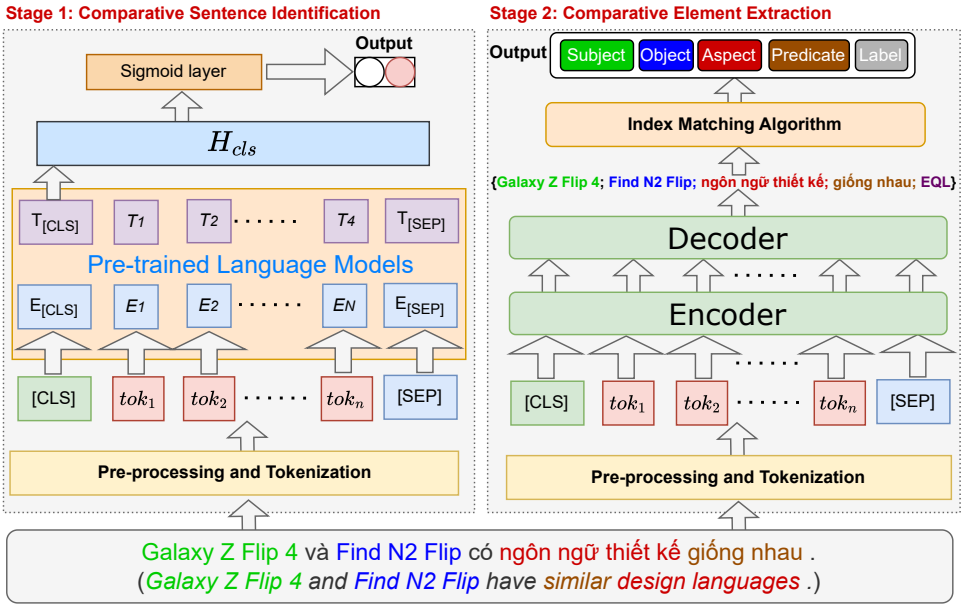


Figure 1: Overall architecture of our two-stages approach for the COQE task

each individually and then combining the results. However, this approach has limitations, as errors in earlier steps can propagate and affect later predictions, reducing the overall accuracy. Additionally, these models fail to consider the relationships between elements within the comparative quintuples mentioned in reviews. Another challenge is the limitation of training data for comparative reviews, which restricts the performance of these models.

To address challenges in the quintuple extraction task, we propose a model that leverages multi-task instruction prompting with conditional generative language models to extract comparative quintuples. The proposed approach combines prompt-based instructions with input reviews, offering the model with richer task-specific context. Additionally, multi-task learning is employed to capture inter-element relationships within quintuples, enhancing performance on the main primary task. To address data limitations, we introduce data augmentation techniques to expand the training dataset.

3. THE PROPOSED FRAMEWORK

The proposed architecture, illustrated in Figure 1, tackles the COQE task. It consists of two stages: identifying whether a review is comparative and then extracting the elements in the quintuplets from the comparative reviews. The details of each stage in our framework are presented below.

3.1. Stage 1: Comparative sentence identification

To determine whether a sentence is comparative, we treat this task as a binary text classification problem. To do this, we employ the fine-tuning approach based on different BERTology language models, that support the Vietnamese language. For a given review X with n words represented as $X = \{x_1, x_2, ..., x_n\}$, after incorporating special tokens ([CLS]

and [SEP]) into the review, we feed the input to the pre-trained BERT-based language model to obtain a set of hidden representations $H = \{h_{cls}, h_1, ..., h_n, h_{sep}\}$, where h_i is the representation of i^{th} word. Then, the hidden state h_{cls} corresponding to the CLS token in the last layer is used as the final representation of the input review. This representation is fed directly into a classification layer with a sigmoid activation to calculate the probability that an input belongs to one of two classes

$$\hat{y} = \text{sigmoid}(W \cdot h_{cls} + b), \tag{1}$$

where W represents the weight matrix and b is the bias term of the classification layer. The sigmoid function is commonly used in logistic regression and binary classification tasks. It has several key properties: it is bounded between 0 and 1, differentiable, and produces an "S-shaped" curve. These characteristics make it ideal for transforming a real-valued input into a probability score. Therefore, the sigmoid function is employed to predict whether the hidden state h_{cls}, derived from the given review, belongs to the "comparative review" class. If the output of the sigmoid function is greater than or equal to a threshold (in our case, 0.5), the review is classified as belonging to the comparative class. To optimize model weights, we use the Weighted Binary Cross-Entropy (BCE) to address data imbalance because the comparative samples are much less than the non-comparative samples

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}\left[w_p \cdot y_i \cdot \log(\hat{y}_i) + w_n \cdot (1 - y_i) \cdot \log(1 - \hat{y}_i)\right], \tag{2}$$

where N is the number of samples, y_i is the true label of i^{th} sample with a value of 0 or 1, $\hat{y}_i$ is the predicted label of the comparative review calculated using Formula (1), w_p and w_n are the weights for the comparative and non-comparative classes. The class weights are calculated from the training set. As noted in the work of Dang et al. [12], there are many pre-trained language models trained on various pre-training corpus and different architectures for the Vietnamese language. Therefore, the performance of models might differ depending on the tasks and domains of the dataset. As a result, we implement our approach using different open-source pre-trained language models, including the monolingual (viBERT [13], vELECTRA [13], PhoBERT [14], and ViDeBERTa [15]) and multilingual (mBERT [16], XLM-R [17], InfoXLM [18], XLM-Align [19], mDeBERTaV3 [20]) versions.

3.2. Stage 2: Comparative element extraction

The purpose of this component is to extract all comparative elements including the subject, object, aspect, predicate, and comparative label in a quintuplet. To accomplish this, we transform it into a conditional text generation problem to utilize the power of pre-trained generative language models. Moreover, we also apply the descriptive instruction prompt tuning strategy with multi-task learning to improve the overall performance of the main task. Specifically, we convert the corresponding outputs of the CEE task and its variant sub-tasks to natural language sentences by designing a generation paradigm. To achieve that, we apply the extraction style modelling [21] to transform the output to a natural sentence as illustrated in the following example: vietnamese

> **Input**: Galaxy Z Flip 4 và Find N2 Flip có ngôn ng thit k ging nhau.
> (*The Galaxy Z Flip 4 and Find N2 Flip have similar design languages*).

Table 1: The list of instruction prompts for the selected tasks

vietnamese

Instruction Prompt
Hãy rút trích b năm thông tin gm ch th, i tng, khía cnh, v t so sánh, loi so sánh trong câu: {Review} (*Please extract the quintuplets, including the subject, object, aspect, predicate, type of comparison in the sentence:* {*Review*})
Hãy rút trích b bn thông tin gm ch th, i tng, khía cnh and v t so sánh trong câu: {Review} (*Please extract the quadruplets, including the subject, object, aspect, predicate in the sentence:* {*Review*})
Hãy rút trích b bn thông tin gm ch th, i tng, khía cnh, loi so sánh trong câu:: {Review} (*Please extract the quadruplets, including the subject, object, aspect, type of comparison in the sentence:* {*Review*})
Hãy rút trích b ba thông tin gm ch th, i tng, khía cnh trong câu: {Review} (*Please extract the triplets, including the subject, object, aspect in the sentence:* {*Review*})
Hãy rút trích b ba thông tin gm ch th, i tng, loi so sánh trong câu: {Review} (*Please extract the triplets, including the subject, object, aspect, type of comparison in the sentence:* {*Review*})
Hãy rút trích b ba thông tin gm ch th, i tng, v t so sánh trong câu: {Review} (*Please extract the triplets, including the subject, object, predicate in the sentence:* {*Review*})
Hãy rút trích b ba thông tin gm khía cnh, v t so sánh, loi so sánh trong câu: {Review} (*Please extract the triplets, including the aspect, type of comparison in the sentence:* {*Review*})
Hãy rút trích b hai thông tin gm ch th, i tng trong câu: {Review} (*Please extract the doublets, including the subject, object in the sentence:* {*Review*})
Hãy rút trích b hai thông tin gm khía cnh, v t so sánh trong câu: {Review} (*Please extract the doublets, including the aspect, predicate in the sentence:* {*Review*})
Hãy rút trích b hai thông tin gm v t so sánh, loi so sánh trong câu: {Review} (*Please extract the doublets, including the predicate, type of comparison in the sentence:* {*Review*})

> **Ground Truth**: {"***subject***": ["1&&Galaxy", "2&&Z", "3&&Flip", "4&&4"], "***object***": ["6&&Find", "7&&N2", "8&&Flip"], "***aspect***": ["10&&ngôn", "11&&ng", "12&&thit", "13&&k"], "***predicate***": ["14&&ging", "15&&nhau"], "***label***": "EQL"}.
> **Extraction Template**: {Galaxy Z Flip 4; Find N2 Flip; ngôn ng thit k; ging nhau; EQL}.

english If the value of the element is empty, it will be replaced with the "Null" value. In case the input review is assigned, multiple quintuplets will be concatenated using a semicolon (;) in the extraction template output. As shown in the above example, we exclude word index information from the output template to reduce the complexity of the model. The corresponding indices of values in the quintuple will be added through an index-matching algorithm. This algorithm is developed based on the fuzzy string matching algorithm combined with heuristic rules based on the analysis of the prediction on the development set.

The purpose of the CEE task is to extract the elements in a quintuple; therefore, we hypothesize that there is correlative information between them. To test this hypothesis, we train our architecture in a multi-task manner on the CEE task and the sub-tasks. We focus on sub-tasks where elements have a correlation relationship rather than considering all possible combinations of problems involving all five elements. This helps the model utilize the knowledge of sub-tasks as well as increase the data for the main task, helping to improve model performance. Furthermore, we also apply diverse descriptive prompts to inform the model of the specific task and facilitate the identification of relationships between different tasks. We notice that using the instruction prompt helps our approach leverage the Knowledge and improves the generalization of the generative models [22]. Hence, we develop the instruction prompt in Vietnamese for the CEE task and corresponding its sub-tasks as shown in Table 1.

english

In this work, we implement an encoder-decoder transformer-based architecture to address

the tasks. We initialize the parameters using various pre-trained language models, including mT5 [23], mT0 [24], viT5 [25] or BARTpho [26]. The list of pre-trained sequence-to-sequence language models used as our main backbones is briefly described below.

- **mT5** [23]: The mT5 models are powerful multilingual versions of the T5 architecture as described in [27], are trained on a massive dataset covering 101 languages. These models utilize the "text-to-text" framework, where each task is conceptualized as transforming input text into output text.
- **mT0** [24]: is a multitask prompt-based fine-tuning variant of the mT5 model [23]. This model was fine-tuned on a diverse cross-lingual task mixture, ensuring robustness and generalizability across various NLP downstream tasks. Additionally, this model is able to enhance performance and improve knowledge transfer across languages.
- **BARTpho** [26]: BARTpho was introduced as a pioneering monolingual sequence-to-sequence model designed to Vietnamese language. However, its pre-training on only 20GB of uncompressed text may limit the performance of BARTpho compared to others trained on larger corpora.
- **viT5** [25]: ViT5 is a SoTA pretrained encoder-decoder language model specifically designed for the Vietnamese language. Built on the architecture of T5 [27], it was trained on a massive dataset of high-quality and diverse Vietnamese texts filtered from the CC100 dataset.

Since the above pre-trained language models utilize different pre-processing techniques on their training data, we adopt the same pre-processing steps for each variant of the aforementioned models. This helps standardize the input data and potentially improve the understanding of models.

3.3. Data augmentation

Throughout the process of analyzing and evaluating the performance of the model, we revealed that substituting or inverting sentence elements (subject or object) in a review can lead to incorrect predictions, even if those elements appeared in the training data. This highlights the sensitivity of models to word order. To address this challenge, we implement a data augmentation strategy to expand training data. This technique replicates existing data, effectively increasing the information available for the model to learn from and mitigating the impact of missing data. Our data augmentation strategy follows these steps:

- **Word set creation**: First, we collect all not-null values for each element in the quintuple to create a specific word set for each element.
- **Sentence and quintuple generation**: For each comparative review, we randomly replace the subject, object, and aspect with elements from their respective word set, generating a new sentence and its corresponding quintuple. Since comparison labels depend on the predicate value, we randomly swap them together to ensure consistency.
- **Data balancing and cleaning**: Finally, we balance the augmented dataset based on the number of comparative elements. We also filter out duplicate samples using sentence similarity with a threshold (e.g. 0.8) and resolve any inconsistencies between the augmented quintuplets and their corresponding reviews (e.g., missing order).

4. EXPERIMENTS

4.1. Dataset and evaluation metrics

In this paper, we use the VCOM dataset [1] collected from the Comparative Opinion Mining from Vietnamese Product Reviews shared task at VLSP 2023. This competition dataset comprises three subsets: the training set, the development and the testing set. Following the previous studies [1, 4], we report the performance for the CSI, CEE and COQE tasks. For the CSI task, we evaluate the Precision, Recall and F1-score using both macro-average and binary-average settings. For the CEE and COQE tasks, we calculate the Precision, Recall and F1-score as the main metrics in the Exact and Binary Match strategies. We also report the results of models in the context of a Five-Element Tuple (T5) [1], based on strategies of Exact and Binary matching.

4.2. Experimental settings

For Stage 1 of the CSI task, we fine-tuned various BERT-based models from Hugging Face using the AdamW optimizer with a batch size of 32 for 10 epochs and a dropout rate of 0.1. Learning rates were set to 2e-5 for large models and 5e-5 for smaller ones. In Stage 2, we fine-tuned generative models with learning rates of 3e-4 for multilingual models (mT5, mT0) and 2e-5 for monolingual models (viT5, BARTPho) for 20 epochs, using batch sizes of 16 or 8. All models used sequence lengths of 256 tokens and were trained on a single NVIDIA A100 80G GPU.

4.3. Comparison systems

In order to show the effectiveness of our framework, we compare it with the following existing approaches for the main COQE task:

- **BERT-LSTM-CRF** [1] consists of two distinct phrases: extraction and validation. In the extraction phase, the input review is fed into BERT-LSTM-CRF to extract the predicted elements. Subsequently, for each extracted predicate value, three separate BERT-LSTM-CRF networks are applied to extract the subject, object, and aspect elements. In the validation stage, the representations of all elements are concatenated and fed into a softmax classifier to verify if the extracted quintuplet is valid. Finally, another softmax layer is used to predict the comparison labels.
- **BERT2BERT** [1] is an encoder-decoder model based on a BERT architecture that generates entire quintuplets for a given input review. If the review is non-comparative, the model outputs a special sequence designated as "n/a". Two post-processing steps are applied to extract the index information and validate the quintuplet from the generated output.
- **Three-Stage BERTs** [11] follows a three-stage process for three sub-tasks: (1) Comparative Sentence Identification, (2) Comparative Element Extraction, and (3) Comparison Type Classification. The first stage uses the PhoBERT [14] to identify the comparative review, while the second stage is solved by ensemble strategy on BERT-based models. Finally, in the third stage the list of possible quadruple combinations in the second stage is fed to the multi-task classification model to detect the comparison labels.

Table 2: Experimental results of BERTology models for CSI task on the development set

Type	Model	Version	Macro-average			Binary-average		
			Precision	Recall	F1-score	Precision	Recall	F1-score
Multilingual Models	mBERT (cased)	base	87.78	88.66	88.21	80.00	82.52	81.24
	mBERT (uncased)	base	88.72	89.95	89.32	81.32	84.81	83.03
	XLM-R	base	88.12	92.66	90.11	78.43	91.69	84.54
		large	89.07	91.98	90.50	80.25	91.98	85.71
	XLM-Align	base	87.45	92.41	89.59	77.11	91.69	83.77
	mDeBERTaV3	base	88.98	91.02	89.95	81.28	87.11	84.09
	InfoXLM	base	85.46	92.51	88.23	72.57	93.98	81.90
		large	88.75	92.77	90.55	79.75	91.40	85.18
Monolingual Models	viBERT	base	86.42	88.69	87.48	77.04	83.67	80.22
	vELECTRA	base	85.50	88.01	86.66	77.07	82.81	79.83
	ViDeBERTa	small	82.34	87.10	84.33	71.14	80.52	75.54
		base	81.96	87.92	84.31	67.64	86.25	75.82
	PhoBERTv2	base	90.18	92.53	91.29	83.02	89.68	86.23
	PhoBERTv1	base	88.87	93.06	90.73	79.85	91.98	85.49
		large	**90.48**	**93.50**	**91.88**	**83.12**	**91.69**	**87.19**

- **Multi-Stage Framework** [4] is designed for the COQE task in English and Chinese. This two-stage framework first extracts the comparative sentence and elements simultaneously. The second stage then verifies the validity of the extracted quadruplets and classifies the comparison type label.

- **Seq2Seq T5** is developed based on T5 model [27] rather than encoder-decoder BERT-based model [1]. This model outputs data in XML format because XML is concise and easily parsed by machines. The large version of T5 [27] is utilized as the main backbone model.

- **Ensemble Framework** is a multi-stage approach based on the architecture proposed by [4], but it differs in the base models for each stage. The first stage leverages an ensemble of a Graph Neural Network and viT5 to identify comparative sentences and extract the first four elements of the quintuplets. The second stage then employs a PhoBERT classifier to predict the comparison type class based on the combination of the input sentence, aspect, and predicate.

5. RESULT AND ANALYSIS

5.1. Main results

As shown in Table 2, the PhoBERTv1 large model achieves the highest overall performance across both multilingual and monolingual categories in terms of evaluation scores. We also observe a significant performance improvement when transitioning from base to large versions of models (e.g., PhoBERTv1, XLM-R, Info-XLM). This highlights the importance of model size in capturing language-specific features. However, large models generally perform better but require more computational resources. Among the base models, PhoBERTv2 achieves the highest performance due to its training on a large-scale monolingual Vietnamese word-segmentation corpus. This targeted training enables it to capture the characteristics of the Vietnamese language more effectively, leading to superior performance on this task.

Table 3: Results of models for CEE Task (Exact Match) on the development set

Models	Version	CEE Task					
		Subject	Object	Aspect	Predicate	Micro F1	Macro F1
mT5	small	60.76	63.42	47.15	53.42	55.78	56.37
	base	66.67	64.59	52.19	55.59	59.08	59.95
	large	67.98	66.79	51.80	59.65	61.47	62.24
mT0	small	63.20	59.57	50.07	52.64	55.34	56.19
	base	65.91	66.54	52.07	55.26	58.98	59.76
	large	68.91	67.55	53.27	59.24	60.89	61.55
BARTpho-syllable	base	66.10	64.72	49.02	55.04	57.88	58.72
	large	**70.59**	69.07	55.02	55.68	61.82	62.59
BARTpho-word	base	69.10	67.51	55.51	60.15	62.45	63.07
	large	68.75	69.02	55.83	**61.76**	63.19	63.84
viT5	base	66.76	67.80	48.57	56.35	59.08	59.87
	large	69.41	**70.04**	**56.41**	61.27	**63.59**	**64.28**

Table 4: Results of different models for the COQE task on the development set

Models	Version	Exact Match		Binary Match	
		Macro F1	Micro F1	Macro F1	Micro F1
mT5	small	1069	2045	1874	35.22
	base	12.66	24.52	24.40	39.74
	large	15.49	27.40	25.55	44.63
mT0	small	10.43	20.26	20.73	35.91
	base	12.89	24.57	24.28	42.74
	large	15.88	26.94	27.46	46.24
BARTpho-syllable	base	12.32	21.17	22.63	37.97
	large	17.15	28.66	**32.16**	**46.58**
BARTpho-word	base	14.18	25.62	23.83	42.03
	large	15.35	27.65	25.62	44.49
viT5	base	16.16	28.47	24.51	42.86
	large	**18.43**	**30.29**	27.35	45.19

In the CEE task, the monolingual viT5 model achieves the highest performance on both aggregated metrics (micro and macro F1), closely followed by the large version of BARTpho-word. This proves that monolingual models trained on large Vietnamese datasets give better results than multilingual models. Comparing the results of mT5 against mT0 shows that there is no significant difference between their results. In fact, the original mT5 model performs slightly better than mT0 in terms of micro and macro F1 scores. As shown in Table 3, the performance of the "subject" and "object" elements is higher than that of the other two parts in all models. Meanwhile, the models yield the lowest results for the "Aspect" elements. One reason for the poor performance of "Aspect" is the diversity of vocabulary and expressions used to express comparative aspects.

Table 4 presents the performance of various generative models for the COQE task, measured by Exact Match (EM) and Binary Match (BM). The viT5 model (large version) outperforms other models in terms of EM F1-score. The highest performance on BM F1-scores is achieved by the BARTpho-syllable with a macro F1 of 32.16% and a micro F1 of 46.58%. It can be observed that the performances of all models show significant improvements under the binary match strategy. Among the three tasks, CSI and CEE achieved promising per-

Table 5: Performance comparison for the COQE task on the test set

Approach	Macro Average			Micro Average		
	Precision	Recall	F1-score	Precision	Recall	F1-score
BERT-LSTM-CRF [1]	5.25	9.98	6.68	9.82	20.93	13.37
BERT2BERT [1]	10.41	8.46	9.23	20.26	16.85	18.40
Three-Stage BERTs [11]	9.68	10.65	9.97	16.75	18.94	17.78
Multi-Stage Framework [4]	9.64	13.75	11.19	17.09	26.98	20.92
Seq2Seq T5	20.93	21.99	21.31	**29.41**	29.41	29.41
Ensemble Framework (GNN,viT5,PhoBERT)	20.21	27.18	23.00	22.34	**33.59**	26.84
Our system (viT5$_{large}$)	**22.16**	**28.62**	**23.73**	28.80	30.29	**29.52**

Table 6: Results of different variations based on our framework on the test set

Approach	Macro Average			Micro Average		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Single viT5	17.57	15.79	16.61	27.39	25.22	26.26
Single viT5 + DA	15.95	14.41	15.10	28.15	26.10	27.09
Single viT5 + Prompt	17.21	16.03	16.59	29.86	28.85	29.34
Single viT5 + Prompt + DA	17.15	16.38	16.72	**30.09**	28.30	29.17
Multi-task viT5	20.55	15.63	17.64	27.58	21.81	24.35
Multi-task viT5 + DA	21.11	16.10	18.17	27.96	22.36	24.85
Multi-task viT5 + Prompt	24.17	21.80	22.51	28.13	29.41	28.76
Multi-task viT5 + Prompt + DA	**28.62**	**22.16**	**23.73**	28.80	**30.29**	**29.52**

formance, while the COQE task exhibited poor performance. This suggests that COQE is the most difficult and challenging task.

Table 5 shows the macro and micro F1-scores of our approach based on viT5 (large version) comparing other approaches on the test set. It can be seen that the BERT-LSTM-CRF and BERT2BERT models achieved the lowest macro F1 scores among the baseline approaches. This may be due to limitations in these architectures for complex tasks like COQE. BERT-based models combined with LSTM and CRF may not be powerful enough to capture the nuances required for this task, potentially leading to poorer performance. For multi-stage approaches like Three-Stage BERTs [11] and Multi-Stage Framework [4] yields better performance, but the difference is not significant. Interestingly, Seq2Seq T5 achieves a performance comparable to our approach on the micro-average F1-score. However, our method outperforms Seq2Seq T5 by around 2% on the macro-average F1-score. Also, we observe that the Ensemble Framework is quite competitive with our system in the macro F1-score and achieves the highest micro average recall. However, this approach requires significantly more computational resources to train and execute the entire system, including the base models used for the voting ensemble.

Table 6 presents the results of different variants of models for the COQE task. It can be seen that the instruction prompt significantly improves the overall performance, particularly for the multi-tasking model, compared to models without integrated prompts. Besides, the data augmentation technique also enhances the performance of micro-average F1 scores, but the improvement is not statistically significant. This limitation arises because these techniques only augment samples based on existing elements within the training set. Another finding is that training the model in a multi-task manner leverages the correlation information between comparison elements, resulting in improved efficiency compared to a single-task model, especially for the macro-average scores.

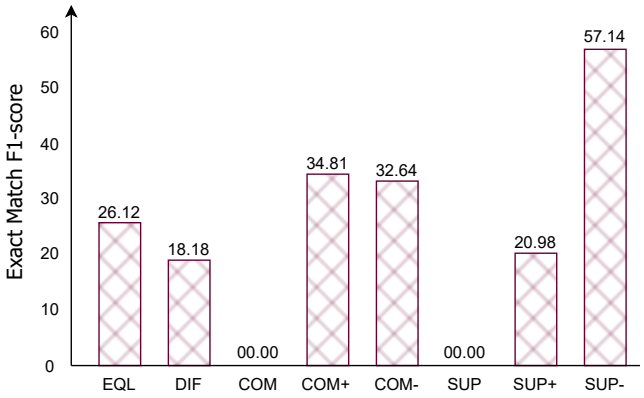
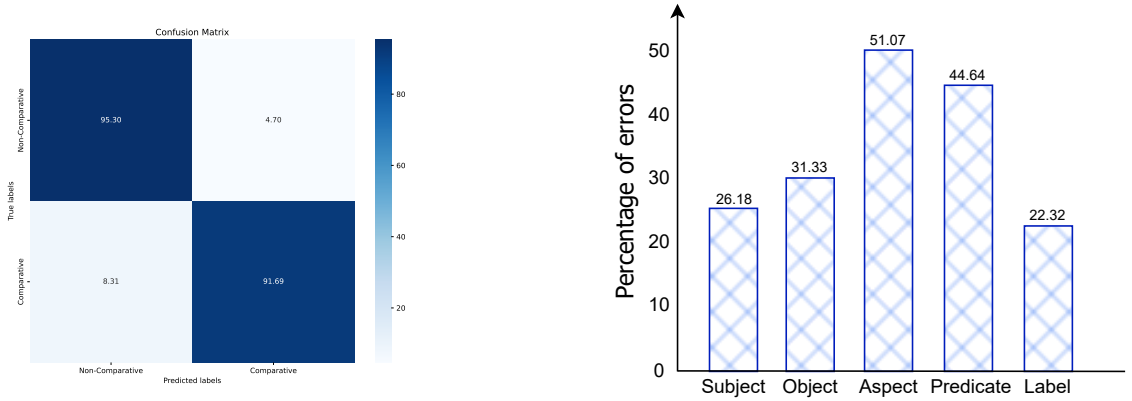


Figure 2: Exact Match F_1-score for Comparison Type labels



(a) Confusion matrix of the Comparative Sentence Identification task

(b) The percentage of errors for five elements in the quintuplet

Figure 3: The confusion matrix and the percentage of errors for five elements

Figure 2 presents the F1 score of the quintuplet under the exact match strategy for the comparison type label. It can be seen that our system is not able to predict the exact quintuplets for the "SUP" and "COM" comparison type labels. One of the reasons for this result is that these labels have one of the lowest numbers of labelled samples in the dataset (as refer in previous work [1]). Interestingly, the prediction on the "SUP-" label achieves the highest performance despite having only 11 samples in the entire dataset. To investigate the high results for the "SUP-" label, we examined the data and found that the predicted samples are too similar in the training set; therefore, the model achieved high performance on the 'SUP-" label. However, this is likely due to overfitting and may not work well with other patterns for this label.

5.2. Error analysis and discussion

english

Figure 3a presents the performance of our model for identifying the comparative review

Table 7: Examples of Review, Ground Truth, and Predicted Labels for CSI task vietnamese

Review	Ground truth	Prediction
alt : Màn hình rng mang li tri nghim s dng tt nht (*alt: Wide screen provides the best user experience*)	Comparative	Non-comparative
Mi in thoi u có nhng tính năng c bit riêng (*Each phone has its own special features*)	Comparative	Non-comparative
Camera ca Galaxy A31 tt hn A12 (*The camera of the Galaxy A31 is better than the A12*)	Non-comparative	Comparative
S khác bit gia hai máy cách b trí cm camera sau (*The difference between the two devices is in the layout of the rear camera cluster*)	Non-comparative	Comparative
C hai in thoi u c h tr sc nhanh (*Both phones support fast charging*)	Non-comparative	Comparative

Table 8: Percentage of error categories for the first four elements in quintuplets

Error Category	Subject	Object	Aspect	Predicate
Indexing Issue (Error 1)	3.28	2.74	-	-
False Positive Error (Error 2)	16.39	17.81	16.81	-
False Negative Error (Error 3)	8.20	16.44	12.61	-
Incomplete Prediction Error (Error 4)	19.67	17.81	13.45	42.31
Over-prediction Error (Error 5)	27.81	20.55	16.81	23.03
Irrelevant Prediction Error (Error 6)	27.87	27.40	40.34	27.88
Partial Match Error (Error 7)	-	-	-	6.73

or not. For non-comparative reviews, the model correctly identified 95.30% of cases but misclassified 4.70% as comparative reviews. For comparative reviews, the model's accuracy was slightly lower, correctly identifying 91.69% of cases and misclassifying 8.31% as non-comparative. After analyzing the predictions, we found that most errors occurred in reviews lacking explicit subjects or objects. Additionally, we identified some misannotated reviews during the data annotation process. Table 7 shows some examples with ground truth labels and corresponding predictions.

In addition, we analyze the performance of our method in extracting comparative elements within correctly predicted comparative reviews. We observed that 25 reviews were predicted to have more quintuples, and 44 reviews were predicted to have fewer quintuples than the ground truth. After filtering out the correctly identified quintuplets, we analyze sets of incorrectly predicted quintuplets to identify the types of errors that occur in different elements. Figure 3b presents the percentage of errors corresponding to elements in the quintuplet. It is obvious that the values of "aspect" and "predicate" have the highest misprediction rates among the five elements.

To gain a deeper understanding of the specific error types for each element, we analyze and categorize the errors in our incorrectly predicted quintuplets. Table 8 presents the statistics of seven error categories corresponding to each element. The first error is related to the index matching algorithm, where there are 3.28% and 2.74% error rates for subject and object elements. This likely occurs because the elements appear multiple times in close proximity within the comment. This proximity might confuse the index-matching algorithm, leading it to assign incorrect positions to the subject and object value.

The second error category is the false positive error, where the ground truth has no

value for an element, but the model predicts a value. This error only appears for the first three elements, including subject, object and aspect, with the same rates of 16.39%, 17.81% and 16.81%. Similarly, the third error category is the false negative error, where the ground truth has a value for an element, but the model predicts it as empty. The error rate for the "subject" and "aspect" elements shows a decreasing trend, but for the "object" element, the reduction compared to error type 2 is not significant. These errors typically occur with quintuplets when one of the two expected elements, subject or object, is assigned the value None. This indicates the model is having difficulty determining the correct values for subject and object. As a result, the current method appears to struggle with handling implicit elements. Additionally, the model struggles to identify complex comments containing multiple quintuplets. The scarcity of training data for reviews containing multiple quintuplets may be the reason for this result.

The fourth error type we encountered is the Incomplete Prediction Error. This error occurs when the model only extracts a portion of the value of the element compared to the ground truth. As shown in Table 8, the "predicate" element has the highest error rate (42.31%) for error type 4, followed by "subject" (19.67%), "object" (17.81%), and "aspect" (13.45%). The high error rate for "predicate" elements is due to the abundance of adverbs expressing degrees of comparison in Vietnamese. This might be challenging for the model to extract fully, especially with limited training samples. Unlike the fourth error (incomplete prediction), the fifth error type occurs when the model predicts additional information for the element, exceeding the ground truth value. In simpler terms, the model makes the correct prediction but includes unnecessary details. Interestingly, this error exhibits the highest rate for the "subject" element, followed by "predicate", "object", and "aspect". The next type of error we observed is the Irrelevant Prediction Error, where the model predicts values unrelated to the ground truth. This error category occurs most frequently for the 'aspect' category, with a percentage of 40.34%. We observed that reviews with many quintuplets or implicit aspects are often misidentified during the prediction process. The final error we observed occurred when the predicted and ground truth labels only partially overlapped. This error only accounts for a small percentage of about 6.73% for the "predicate" element.

6. CONCLUSION AND FUTURE WORK

In this work, we introduce an effective framework for the COQE task. This framework aims to identify comparative sentences from product reviews and extract their corresponding comparative opinion quintuples. We first identify the comparative review based on the fine-tuning pre-trained language models. Then, we propose a novel approach to extract all the quintuplets in a comparison review as a conditional text generation task. To enhance the performance of models, we apply the descriptive prompt tuning strategy with multi-task learning combined with the simple data augmentation technique to improve the overall performance of the COQE task. Experimental results showed that our framework outperforms the other baselines and previous approaches for the Vietnamese VCOM dataset [1]. In future work, further exploration of data augmentation techniques is a promising direction.

ACKNOWLEDGMENT

This research is funded by University of Information Technology-Vietnam National University Ho Chi Minh City grant number D1-2024-21.

REFERENCES

[1] H. Q. Le, C. C. Duy, V. N. Khanh, and V. T. Mai, "Overview of the VLSP 2023 – comom shared task: A data challenge for comparative opinion mining from Vietnamese product reviews," in *The 10th Workshop on the AI-powered Vietnamese Language and Speech Processing*, 2024.

[2] L. Bing, *Sentiment Analysis and Opinion Mining*. Springer Nature Switzerland AG, 2012.

[3] N. Jindal and B. Liu, "Identifying comparative sentences in text documents," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, p. 244–251.

[4] Z. Liu, R. Xia, and J. Yu, "Comparative opinion quintuple extraction from product reviews," in *Empirical Methods in Natural Language Processing*, 2021, pp. 3955–3965.

[5] Z. Yang, F. Xu, J. Yu, and R. Xia, "UniCOQE: Unified comparative opinion quintuple extraction as a set," in *Findings of Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 12 229–12 240.

[6] Q. Xu, Y. Hong, F. Zhao, K. Song, Y. Kang, J. Chen, and G. Zhou, "Low-resource comparative opinion quintuple extraction by data augmentation with prompting," in *Findings of Empirical Methods in Natural Language Processing*, 2023, pp. 3892–3897.

[7] F. Gao, Y. Liu, W. Fu, M. Zhang, A. Ballard, and L. Zhao, "End-to-end comparative opinion quintuple extraction as bipartite set prediction with dynamic structure pruning," *Expert Systems with Applications*, 2023.

[8] J. Nitin and L. Bing, "Mining comparative sentences and relations," in *Association for the Advancement of Artificial Intelligence*, 2006, p. 1331–1336.

[9] Q. Xu, Y. Hong, F. Zhao, K. Song, J. Chen, Y. Kang, and G. Zhou, "GCN-based end-to-end model for comparative opinion quintuple extraction," *International Joint Conference on Neural Networks*, pp. 1–6, 2023.

[10] N. X. Bach, P. D. Van, N. D. Tai, and T. M. Phuong, "Mining Vietnamese comparative sentences for sentiment analysis," in *IEEE International Conference on Knowledge and Systems Engineering*, 2015, pp. 162–167.

[11] L. Ha, B. Tran, P. Le, T. Nguyen, D. Nguyen, N. Pham, and D. Huynh, "Unveiling comparative sentiments in Vietnamese product reviews: A sequential classification framework," *arXiv preprint arXiv:2401.01108*, 2024.

[12] V. T. Dang, D. N. Hao, and L. T. N. Nguyen, "Vietnamese sentiment analysis: An overview and comparative study of fine-tuning pretrained language models," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 6, 2023.

[13] T. V. Bui, T. O. Tran, and P. Le-Hong, "Improving sequence tagging for Vietnamese text using transformer-based neural models," in *Pacific Asia Conference on Language, Information and Computation*, 2020, pp. 13–20.

[14] D. Q. Nguyen and N. A. Tuan, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of Empirical Methods in Natural Language Processing*, 2020, pp. 1037–1042.

[15] C. D. Tran, N. H. Pham, A. T. Nguyen, T. S. Hy, and T. Vu, "ViDeBERTa: A powerful pre-trained language model for Vietnamese," in *Findings of European Chapter of the Association for Computational Linguistics*, 2023, pp. 1071–1078.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *the Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, 2019, pp. 4171–4186.

[17] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451.

[18] Z. Chi, L. Dong, F. Wei, N. Yang, S. Singhal, W. Wang, X. Song, X.-L. Mao, H. Huang, and M. Zhou, "InfoXLM: An information-theoretic framework for cross-lingual language model pre-training," in *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, 2021, pp. 3576–3588.

[19] Z. Chi, L. Dong, B. Zheng, S. Huang, X.-L. Mao, H. Huang, and F. Wei, "Improving pretrained cross-lingual language models via self-labeled word alignment," in *Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 2021, pp. 3418–3430.

[20] P. He, J. Gao, and W. Chen, "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing," in *International Conference on Learning Representations*, 2022.

[21] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "Towards generative aspect-based sentiment analysis," in *Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 2021, pp. 504–510.

[22] V. Sanh, A. Webson, C. Raffel, S. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, A. Raja, M. Dey *et al.*, "Multitask prompted training enables zero-shot task generalization," in *International Conference on Learning Representations*, 2021.

[23] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, 2021, pp. 483–498.

[24] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, and et al., "Crosslingual generalization through multitask finetuning," in *Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 15 991–16 111.

[25] L. Phan, H. Tran, H. Nguyen, and T. H. Trinh, "ViT5: Pretrained text-to-text transformer for Vietnamese language generation," in *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, 2022, pp. 136–142.

[26] N. L. Tran, D. M. Le, and D. Q. Nguyen, "BARTpho: Pre-trained sequence-to-sequence models for Vietnamese," in *INTERSPEECH*, 2022.

[27] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

Received on April 19, 2024
Accepted on January 04, 2025